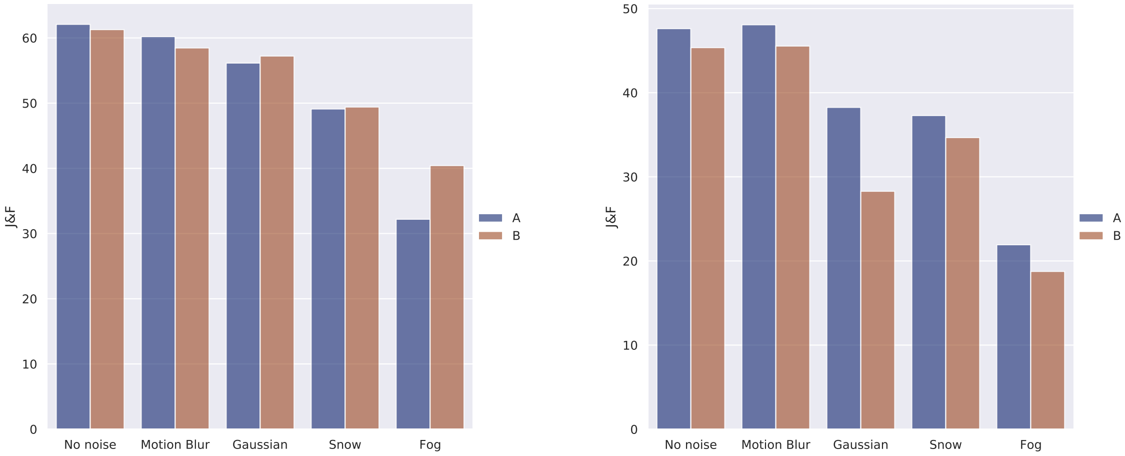
Supplementary Materials: Self-supervised Test-time Adaptation on Video Data

TAO-VOS Subset
YFCC100M/v_a0218442a084abd3c292822fc1d6bfb
YFCC100M/v_8ddabfb1e0eea9fc468448e936378b
LaSOT/airplane-3
LaSOT/airplane-11
YFCC100M/v_2a9ff118df23ac28f710d283fab1fe81
Charades/9M48H
HACS/Clipping_cat_claws_v_ogus-Ik3UMA
HACS/Playing_squash_v_IUbQjSiZL-Y_scene_0_171-2322
HACS/Raking_leaves_v_iSjk42F0rvM
HACS/Raking_leaves_v_KUdBvuRaAbk_scene_0_0-6975
HACS/Rock-paper-scissors_v_LFPYYYZstjg_scene_0_0-1790
LaSOT/airplane-4
LaSOT/bicycle-18
LaSOT/bird-12
LaSOT/bird-17
LaSOT/cat-4
LaSOT/cat-7
LaSOT/cattle-9
LaSOT/deer-15
LaSOT/deer-4
LaSOT/hat-18
LaSOT/helmet-1
LaSOT/lion-9
LaSOT/lizard-16
LaSOT/rabbit-12
LaSOT/rabbit-17
LaSOT/racing-11
LaSOT/racing-14
LaSOT/shark-18
LaSOT/shark-20
LaSOT/sheep-15
LaSOT/skateboard-1
LaSOT/skateboard-12
LaSOT/skateboard-18
LaSOT/spider-5
LaSOT/surfboard-19
LaSOT/surfboard-8
LaSOT/turtle-16
LaSOT/umbrella-18
LaSOT/zebra-19
YFCC100M/v_bcfdfcfdc8dfd352d18ce4698eb46



(a) VideoWalk [14]

(b) MAST [18]

Figure 3: The performance of TTT [36] on VideoWalk and MAST when **(A)** freezing the normalization statistics during training and then updating with the best-found momentum according to Figure 2 in the last step, while **(B)** corresponds to finetuning when using normalization statistics computed from the target video (as in the standard `train` mode in PyTorch [27]). For MAST, it is better to freeze the statistics at the finetuning stage, while for VideoWalk, we observe that it is sometimes better to train using the normalization statistics of the target domain *e.g.*, for Fog. This observation correlates with the plots in Figure 2 where we can see the performance in VideoWalk is best when completely replacing the normalization statistics with those collected from the target video. From our experiments, we observed that TENT [42] also follows a similar pattern.